

AFFTC-PA-11601



## Monte Carlo Techniques for Estimating Power in Aircraft T&E Tests

Mr. Todd Remund  
Dr. William Kitto

AIR FORCE FLIGHT TEST CENTER  
EDWARDS AFB, CA

16 AUGUST 2011

Approved for public release A: distribution is unlimited.

AIR FORCE FLIGHT TEST CENTER  
EDWARDS AIR FORCE BASE, CALIFORNIA  
AIR FORCE MATERIEL COMMAND  
UNITED STATES AIR FORCE

A  
F  
F  
T  
C

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. <b>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.</b>					
1. REPORT DATE (DD-MM-YYYY) 16-08-2011		2. REPORT TYPE ITEA Conference		3. DATES COVERED (From - To) 20-07-2011 to 22-07-2011	
4. TITLE AND SUBTITLE  Monte Carlo Techniques for Estimating Power in Aircraft T&E Tests				5a. CONTRACT NUMBER NA	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)  Todd Remund, Dr. William Kitto				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AND ADDRESS(ES)  812 TSS 307 E. Popson Ave Edwards AFB, CA 93524				8. PERFORMING ORGANIZATION REPORT NUMBER  AFFTC-PA-11601	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) 812 TSS 307 E. Popson Ave Edwards AFB, CA 93524				10. SPONSOR/MONITOR'S ACRONYM(S) N/A	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S) N/A	
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release A: distribution is unlimited.					
13. SUPPLEMENTARY NOTES CA: Air Force Flight Test Center Edwards AFB CA                      CC: 012100					
14. ABSTRACT Edwards AFB, as a matter of policy, requires statistical rigor be a part of test design and analysis. Statistically defensible methods are used to gain as much information as possible from each test. This requires: <ul style="list-style-type: none"> <li>Statistically defensible methods be identified and applied to each test</li> <li>Setting up tests to maximize scope of inference, and</li> <li>Determining the power or each test to optimize sample size</li> </ul> This paper demonstrates how Monte Carlo techniques may be applied to aircraft test and evaluation to determine the power of the test and the associated sample size requirements. Traditional methods for determining the power of a test are based on distributional assumptions associated with data. These assumptions may not be appropriate; a distribution-free Monte Carlo technique for power assessment for tests with (possible) serially correlated data is presented. The technique is illustrated with an example from a target location error (TLE) test. Power of the test and appropriate sample sizes are derived using Monte Carlo simulation implemented in R.					
15. SUBJECT TERMS Power, statistics, resampling, Monte Carlo simulation, R, sample size, CEP, CE90, circular error.					
16. SECURITY CLASSIFICATION OF: Unclassified			17. LIMITATION OF ABSTRACT  None	18. NUMBER OF PAGES  16	19a. NAME OF RESPONSIBLE PERSON 412 TENG/EN (Tech Pubs)
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER (include area code) 661-277-8615

## **Monte Carlo Techniques for Estimating Power in Aircraft T&E Tests**

**Remund, Todd**

**USAF AFMC 812 TW/EN**

**Edwards AFB, CA**

**[Todd.Remund@Edwards.AF.MIL](mailto:Todd.Remund@Edwards.AF.MIL)**

**Kitto, William**

**USAF AFMC 812 TW/EN**

**Edwards AFB, CA**

**[William.Kitto@Edwards.AF.MIL](mailto:William.Kitto@Edwards.AF.MIL)**

### **ABSTRACT**

Edwards AFB, as a matter of policy, requires statistical rigor be a part of test design and analysis. Statistically defensible methods are used to gain as much information as possible from each test. This requires:

- Statistically defensible methods be identified and applied to each test
- Setting up tests to maximize scope of inference, and
- Determining the power of each test to optimize sample size

This paper demonstrates how Monte Carlo techniques may be applied to aircraft test and evaluation to determine the power of the test and the associated sample size requirements. Traditional methods for determining the power of a test are based on distributional assumptions associated with data. These assumptions may not be appropriate; a distribution-free Monte Carlo technique for power assessment for tests with (possible) serially correlated data is presented. The technique is illustrated with an example from a target location error (TLE) test. Power of the test and appropriate sample sizes are derived using Monte Carlo simulation implemented in R.

**Keywords:** Power, resampling, Monte Carlo simulation, R, sample size, CEP, CE90, circular error.

## **Introduction**

Power is a well documented and well researched statistical tool useful for determining the number of samples to gather for planning a scientific experiment. There are many papers and books written on the topic. Many of the more common problems in statistics such as t-tests, ANOVA, regression, and other linear model methods seem to have a ‘closed’ form solution to power calculation and sample size determination. Despite all these efforts and successes, there is still a wide range of test conditions where statistical power methodology is not readily available. Conventional methods of determining the power of a test assume a distributional form for the statistic of interest is known. In situations this is not the case, power is difficult to estimate.

In test and evaluation (T&E) there is a need to specify the analysis of radial error data. Target location error seems to be one out of many arenas requiring analysis of radial error. There are many methods of calculating upper confidence limits on percentiles of radial error, and comparing these to specifications of one sort or another. When T&E plans a test, there is an inherent question directly related to the cost, “How many data points do I need?” This leads to the need of power calculations that guide in sample size choice. But the method presents a challenging power calculation if parametric methods based on probability distributions are sought.

A new method of calculating confidence limits for radial error has been proposed (Hurwitz et al 2011) and is now used as an example of how Monte Carlo methods can be employed to calculate power for sample size determination. We shall use circular error (CE), i.e. radial error, as our measure and focus on computing statistics with regard to the 90<sup>th</sup> percentile of the CE distribution. Obviously CE can be analyzed for any percentile, and the methodology in this paper allows for analysis of any percentile and confidence level.

## **Statistical Power**

When performing any hypothesis test, or even in day to day decision making, there is always a chance that the decision that was made is wrong. If a change is determined to be necessary, it is best to identify a real change—that is, a change of practical significance. A new software load may, in fact, degrade point accuracy by a fraction of a degree. This typically has no effect on tactical use of the system, so it can be ignored. There are two types of errors and two types of correct decisions. Table 1 depicts the problems that are prevalent in any decision.

Table 1: Matrix depicting correct and incorrect decisions

Truth	Decision	
	$H_0$ : No Change	$H_A$ : Change
$H_0$ : No Change	Correct Negative	False Positive
$H_A$ : Change	False Negative	Correct Positive

Each of the four possible outcomes has a corresponding hypothetical probability. It is obviously best to maximize the ‘Correct’ outcomes subject to cost constraints. In a statistical test, the ‘False Positive’ outcome is fixed at a preselected level, called  $\alpha$ . The ‘Correct Negative’ outcome has probability in complement to  $\alpha$  and is called confidence,  $1-\alpha$ . On the other hand, a ‘False Negative’ is denoted with the Greek letter  $\beta$ , and the corresponding complement is called power. Power is the object of interest. If it is decided that there is reason to declare change it is optimal to have high probability of getting it right when it is practical to do so.

Consider the process of sighting in a rifle. You need a certain number of rounds to decide on a correction to the scope or open sights. In order to do this it is important to realize that there is always going to be some discrepancy in the scope or open sights, and possibly in the rifle as well, that will cause a bias in average hit point location. The real decision is how big of a discrepancy warrants changing the setting on the aiming apparatus? If 100 yards is the distance of interest, is it really important if the rifle is on average 2 inches off? What about 6 inches? This decision determines how many rounds are needed to properly estimate the average hit point and the error surrounding that point, and ultimately amounts to a cost of sighting in the rifle.

Alternately consider the choice of the rifle when making a purchase. Suppose a gun shop allows the would-be buyer to test the rifle previous to buying. Suppose it is desirable that the spread is no larger than a specified radius for 90% of the shots? At the very least, a gun owner would desire to have as precise a rifle configuration as possible given money constraints. This may influence the decision to buy one rifle configuration over another. In other words, the spread of the holes in the target is not controlled by an adjustment of a knob; it is inherent in the way the rifle was made, and the exterior environmental factors that have not been controlled. Suppose the rifle is mounted in a fixed gun stand, and the environmental factors are controlled sufficiently to delineate the spread to hidden factors in the rifle. The objective at this point is to test certain rifles of interest and determine which rifles satisfy the prescribed radial error limitation, or at least find which is closest to the limit. This example parallels the target location error analysis done in many T&E exercises, whether dealing with coordinates from a pod, or actual hit points of a weapon.

By choosing the power and size of the detectible correction, the sample size is automatically determined for the test. Before moving on, it is helpful to give an exact definition of power in mathematical terms.

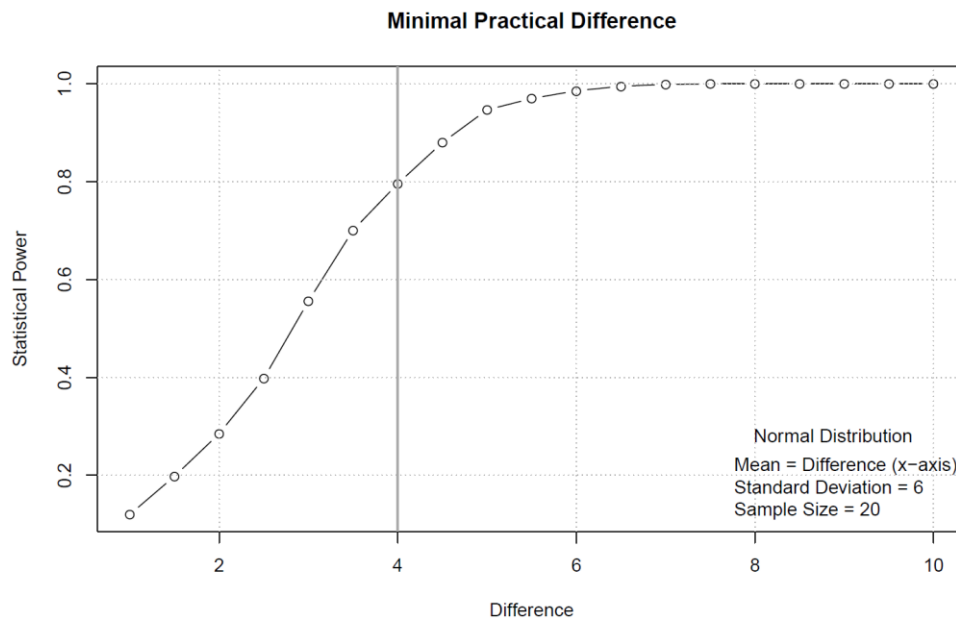
$$\Pr(\text{revert to Alternate} \mid \text{Alternate is the truth}) = \Pr(H_A | H_A) = \text{power}$$

In words this says, “The probability of reverting to the alternate hypothesis given the alternate hypothesis is correct.” Essentially, the probability of saying there is a difference from zero when there really is a difference from zero.

### Minimal Alternate Hypothetical Population

The minimal alternate hypothetical population (MAHP) is a concept derived from statistical hypothesis test philosophy. In statistical hypothesis testing there is a null hypothesis and an alternate hypothesis. Data is gathered for the purpose of testing the null hypothesis to see if it holds. Consider the common one-sample t-test in statistical methodology; we are testing to see if the mean of the data is equal to zero. After having performed a power analysis a sample size is determined that provides the ability to see a certain size difference between zero and a mean that technically is not zero. This difference is labeled  $\delta$  which is the *minimal* difference that researchers would like to see with specified power, or probability of correct detection. The researcher might say, “If the mean of the data is at least 4 feet, I would like to see it?” This minimal difference is the smallest size difference the researcher cares about; everything smaller than this is of no practical consequence.

Figure 1: Practical difference and its effect on power

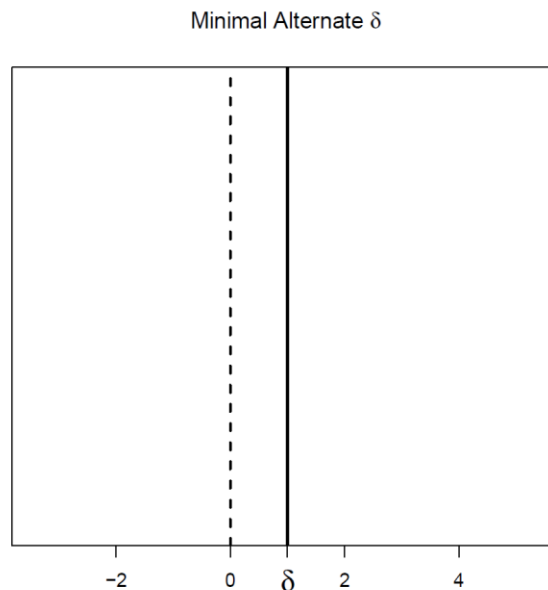


Notice in Figure 1 that if 80% power is desired and the minimal practical difference of 4 is determined, all other size differences of greater magnitude have higher power. So the minimal

difference is the single value that is detectable across 80% of future samples. All larger differences have higher chance of detection, and those of smaller size have a lower chance of detection. Hence the mean of the data, which is a sample from the population, may be larger or smaller than this specified difference—if smaller it is of no consequence and operationally equivalent to zero.

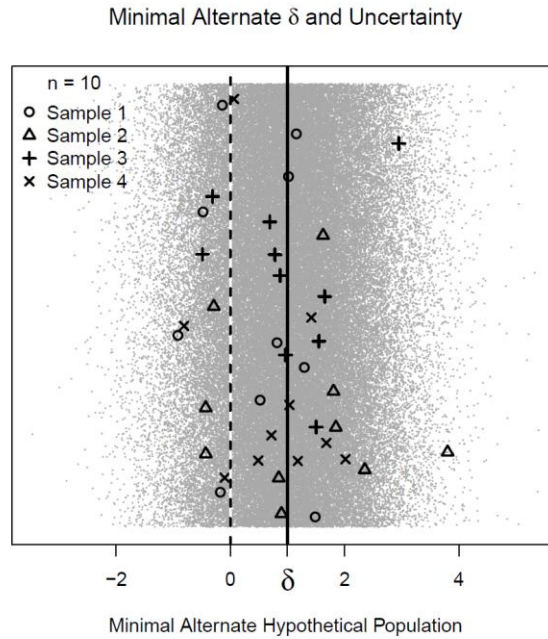
If a difference is detected by the t-test, the null hypothesis is rejected and we revert to the alternate hypothesis. The idea here is to build a hypothetical population that contains the minimal difference inherently; with the case of the one-sample t-test it becomes the mean of a hypothetical normal distribution. This minimal difference is called the minimal alternate because it is the smallest  $\delta$  that causes us to revert to the alternate hypothesis in a statistical test with a specific power.

Figure 2: Minimal difference in the hypothetical population



Specifying  $\delta$  is based on the primary research question and represents the smallest practical difference determined by researchers. As in typical power analysis, the uncertainty of the population must be estimated and integrated into the sample size determination.

Figure 3: MAHP specification and sampling for power calculation



The construction of the MAHP is performed by choosing  $\delta$ , determining a reference distribution—in this case the normal is appropriate—uncertainty is factored in using an estimated standard deviation from prior sampled data from a real phenomenon. The normal distribution with  $\delta$  as the mean and measured standard deviation is the MAHP.

## Method

It is known at this time that there is truly a difference inherent in the MAHP, this is the minimal alternate or minimal practical difference determined by the researcher. Repeated sampling of the hypothetical population can provide a basis for power estimation. We simply pose candidate sample sizes,  $n$ , and repeatedly sample  $n$  values from the MAHP. Generate 100,000, or 1,000,000 values then sample using a random sample routine from these simulated values. This is the Monte Carlo simulation.

For each sample from the MAHP the statistical test of interest is performed and the result is recorded as a zero if the test fails to detect a difference between the sample mean and zero. A one is recorded when the test detects a significant difference from zero. Label this vector  $\mathbf{v}$ . Power is then estimated by summing  $\mathbf{v}$ , and dividing by the number of repeats. If 1,000 repeated samplings of the MAHP were performed the divisor would be 1,000.

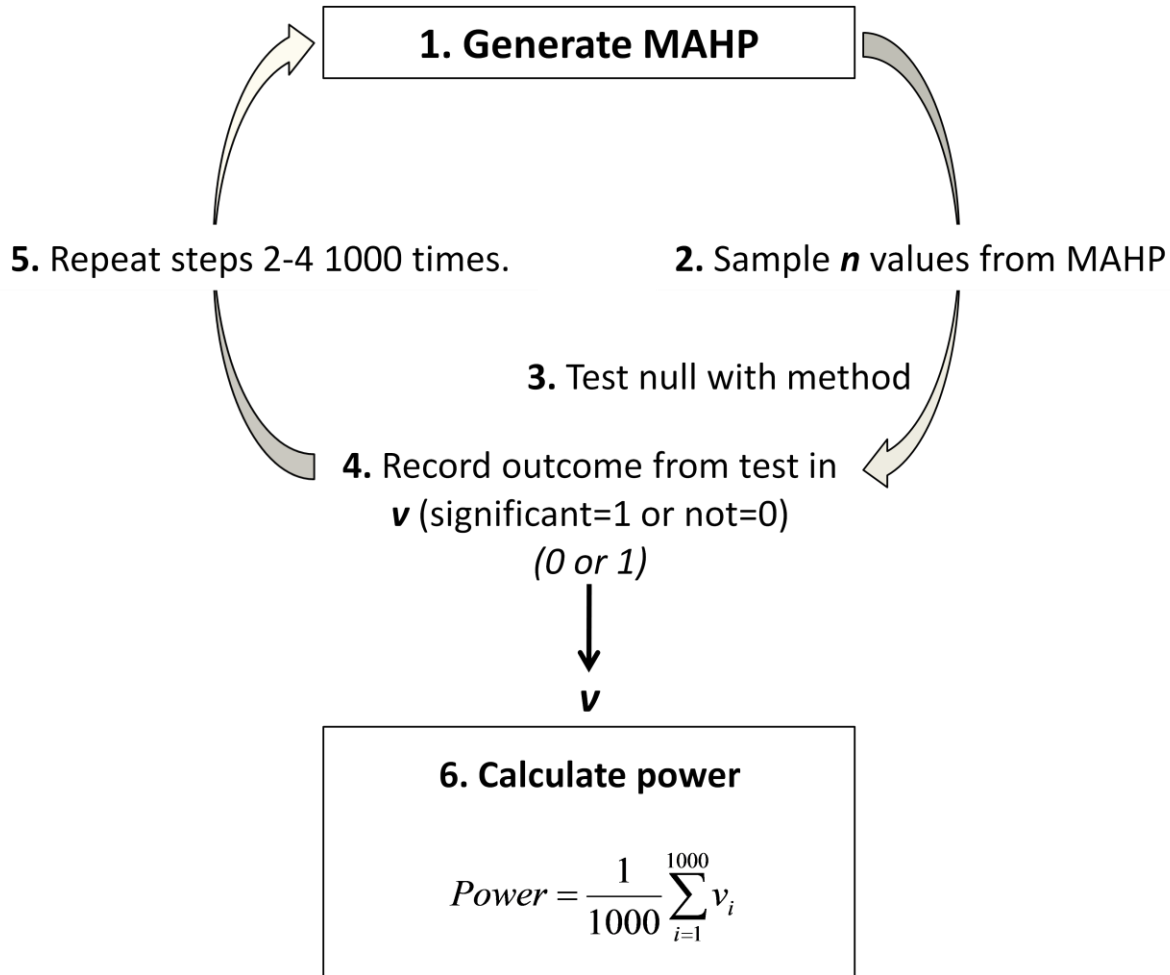
Equation 1

$$power = \frac{1}{1000} \sum_{i=1}^{1,000} v_i$$

The MAHP in Figure 3 is generated with a normal distribution and mean of  $\delta=1$  and standard deviation of  $\sigma=1$ . A candidate sample size of 10 achieved the desired power of 80.2%; the target was 80%. Comparison to conventional power calculation we see a slight deviation. The conventional power value for this situation is 80.31%. The slight discrepancy is due to the nature of Monte Carlo simulation aspect of the method. The estimated value indicates that 802 of the 1,000 repeated samplings produced significant results in the t-tests.

Having stepped through the process of power estimation via Monte Carlo simulation with respect to a one-sample t-test a more general and concise statement for this method is presented in Figure 4.

Figure 4: Monte Carlo power estimation flowchart



### Serially Correlated Data

In this section we take one step toward the unknown. If there are two samples for which the means are to be compared, a t-test may be employed. However, if there is serial correlation present in the data the estimate of standard error of the statistic could be grossly underestimated and the t-test will give misleading results. An adjustment is needed to correct the estimate of standard error.

A simple treatment of serially correlated data is found in (Ramsey & Schafer 2002). Below is a description of the general idea.

Equation 2

$$SE(\bar{x} - \bar{y})_{adjusted} = SE(\bar{x} - \bar{y}) \sqrt{\frac{1 + r_{pl}}{1 - r_{pl}}}$$

Standard error of the statistic is adjusted to account for serial correlation—this is from a simple case where the time series process for the error is an order one autoregressive, AR(1). The estimate of the autocorrelation coefficient,  $r_{pl}$ , is the pooled autocorrelation between x and y, the two samples. Methods to handle more complex serial correlation patterns are found in many statistics books on time series. The test proceeds by computing a confidence interval for the statistic.

Equation 3

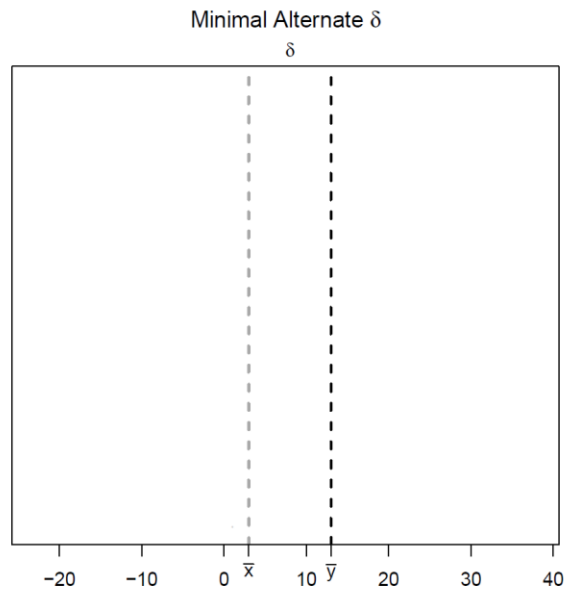
$$CI = \bar{x} - \bar{y} \pm z_{1-\frac{\alpha}{2}} SE(\bar{x} - \bar{y})_{adjusted}$$

The reference distribution for z is the standard normal, mean of zero standard deviation of one. Significance is detected when the interval does not span zero. If the interval does span zero, then there is no evidence of a significant difference. This is another situation where a MAHP can be created and the Monte Carlo power estimate may be used to compute power and sample size. A  $\delta$ , which is the desired minimal difference, is chosen. Standard deviations must be figured, and in addition to these two items, the expected autocorrelation for each group must be found. These are the necessary components for generating the MAHP.

This example will be pursued to reveal that the generation of the MAHP needs to be done with great care. If it is done improperly in this case, the serial correlation built into the hypothetical population will be destroyed and the sample size estimation will be incorrect.

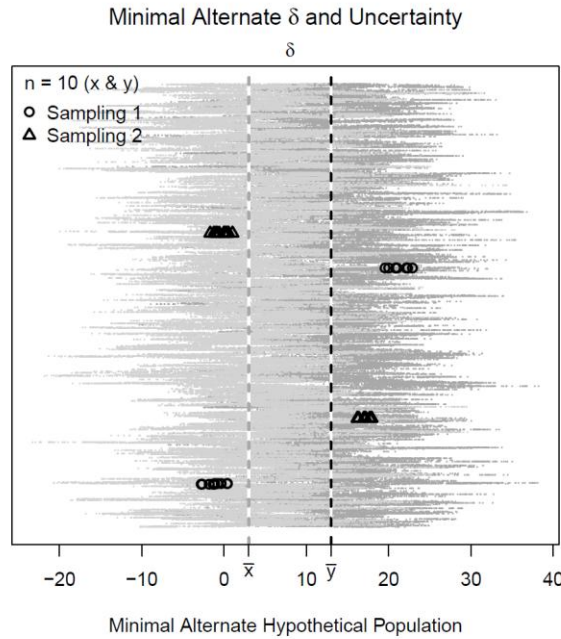
In the previous section the sampling was done by generating a large number of values from the MAHP, then randomly selecting values from the generated values. For serial correlation this does not work. The MAHP can be used to generate the 100,000 values, then starting indexes are randomly drawn. Data is sampled by taking the random start index and taking the value corresponding the start index and the n-1 values following. This preserves the serial correlation built into the MAHP. Set  $\delta$  to a value of 10, meaning the difference in the means is 10 units.

Figure 5: Minimal difference in hypothetical populations, two sample



In Figure 5 the difference between the means is set as defined above, and the location of the differences across the number scale is not necessarily important—it is determined by what is important for the test under consideration. The next step is to incorporate the uncertainty in each group population. Both have the same standard deviation of 1, and the same autocorrelation coefficient of 0.99. Now the distribution can be generated.

Figure 6: MAHP for the two-sample test of means with serial correlation



It is obvious that the random sampling scheme is quite different with a look at Figure 6. In the last example the samples were from random locations in the MAHP. This scenario samples a consecutive string of samples with random starting location. This is proper and necessary to preserve the serial correlation structure. It also represents real life sampling. Any single instance of data measurement will start essentially at a random time point in the history of a process, then the samples following will be linked directly to that initial sample. It is apparent that from the last example to this example the sampling and generation of the MAHP are done differently due to the serial correlation.

Note that the samplings done in this section and also in the previous section could be drawn directly from a normal random number generator. It is necessary to ensure it generates normal distributions with serial correlation, or it must be built in. The above sampling scheme will work given sufficiently large MAHPs, however, it is done in such a way as above to portray the philosophy behind the creation of the MAHP and its use.

$$\text{Equation 4}$$

$$\frac{(v_1 + v_2 + \dots + v_{1000})}{1000} = \frac{(1 + 0 + 1 + \dots + 0 + 1)}{1000} = \frac{862}{1000} = 0.86$$

For a sample size of 10 the power ends up at 86.2%. That is, 862 of the 1,000 repeated samplings from the MAHP produced significant differences. This is using Equation 1 to calculate power by summing up the  $v$  vector containing the indicators of significance and dividing by 1,000. The repeated sampling from the MAHP is done in a statistical analysis program in a loop. The process is again detailed in Figure 4, where the flow of the process is still the same; the MAHP

and statistical test are now different from the previous example. As a side note, if the usual two-sample t-test had been employed with  $\alpha=0.05$ , and the actual confidence would have been in the neighborhood of 24%, and the true  $\alpha$  for that test procedure would be 0.76. That is, we would have claimed a confidence of 95% and been quite mistaken. To remedy this situation the standard error must be adjusted by Equation 2 and the confidence would then be near the desired level.

So far the Monte Carlo power estimate for a one-sample t-test, and a two-sample comparison of means in the presence of serial correlation were presented; the same method works for more complex situations.

### **Circular Error Distributions**

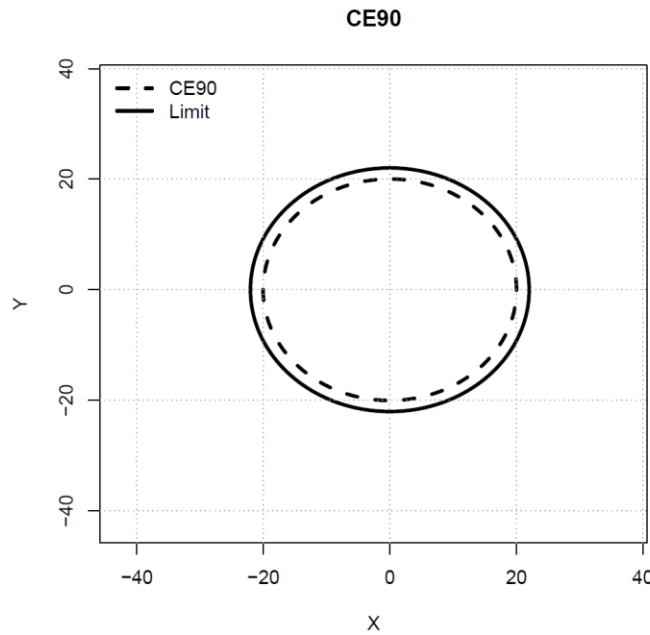
In T&E, targeting devices are evaluated for accuracy and precision. Circular error, particularly percentiles and their upper confidence bounds, answers questions about whether the targeting device is performing adequately or not. Three methods for calculating the upper confidence bound for the percentile of CE were given in (Hurwitz et al 2011). Power for any one of these methods can be calculated using any or all of the methods presented in this paper. The MAHP remains the same for evaluation across any one of these methods. Figure 4 above can be applied, with step number 3 set up for the particular application. Everything else then remains the same.

In this paper the Monte Carlo approach will be used to demonstrate how to find power for a CE90 (90<sup>th</sup> percentile) estimation problem. The first step is to create the MAHP. Suppose there is a limit, i.e. a circular error which must not be exceeded with a targeting device. The CE percentile must not extend beyond this limit. The estimate of CE90 is sometimes used as a measure against the limit to indicate compliance. This is somewhat unhelpful because the value itself is simply a point estimate of the actual population percentile, and hence is subject to sample variability. By using the confidence bound or a statistical test for comparison, we attach a probability statement as to whether we met the limit or not. This takes into account the stochastic nature of the data sample. The upper confidence bound with 95% confidence indicates that the true CE90 (an unknown population parameter) is less than this upper bound 95% of the time across many samplings from the actual population.

With this in mind, we can create the MAHP for CE that puts CE90 within a certain vicinity of the limit; say  $\delta$  feet away from the limit. The estimate of CE90 will be denoted  $\widehat{CE}_{90}$ , and the population parameter value will be labeled CE90. As a hypothetical example, suppose researchers wish to detect a difference between the actual value and the limit if it is no less than 2 feet difference. A specification details the circular error limit to be 22 feet. The MAHP must have a CE90 limit that is detectable with power of 80%. It must be no greater than 20.

So,  $\delta = 2$  feet with the limit set at 22 feet. The mean of  $x$  and  $y$  are set to zero, the standard deviation of  $x$  is  $\sigma_x = 8.87$ , and  $\sigma_y = 8.87$ . This produces a CE90 of 20 feet which is 2 feet below the limit. The correlation between  $x$  and  $y$  is  $\rho = 0.8$  and is used to generate data for the MAHP.

Figure 7: CE90 and the limit

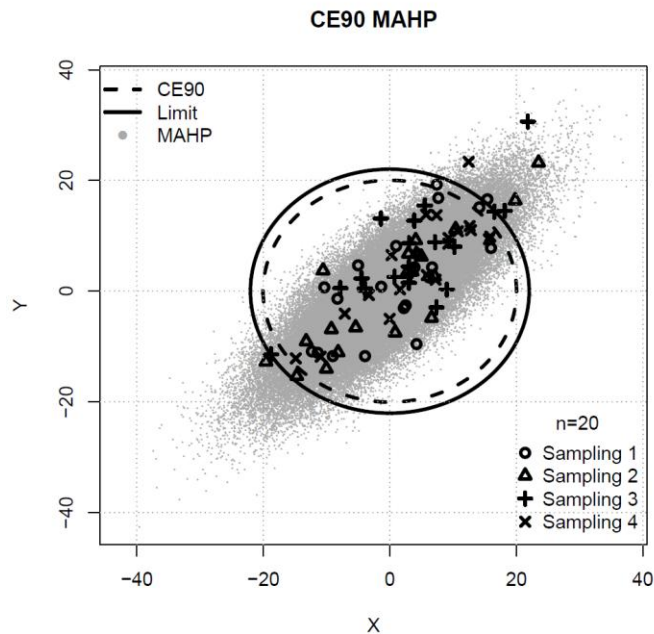


In this situation we need to find the standard deviations and correlation between  $x$  and  $y$  such that the distribution will produce a 90<sup>th</sup> percentile that is 2 feet smaller than the limit. This is a little different from the t-test and two-sample comparison scenarios. The mean and covariance structure in the data determine the  $\delta$ . It becomes apparent that complete knowledge of the distribution, the statistic and the testing method must be had before the MAHP can even be created. As the situation becomes more complex, so does the generation of the MAHP.

Generating points from the MAHP is done by drawing 100,000 samples from a bivariate normal distribution.

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \sigma_x \sigma_y \rho \\ \sigma_x \sigma_y \rho & \sigma_y^2 \end{pmatrix} \right]$$

Figure 8: CE90 MAHP and samplings for power calculation



With a sample size of 300, the power is 76.5%. Again, Equation 1 is utilized in this calculation, after having applied steps 1-5 of Figure 4. The vector of ones and zeros,  $\mathbf{v}$ , is a result of the loop and is applied to the before stated equation. This means that in future samples from the targeting pod we will need 300 samples to ensure that if CE90 is within 2 feet of the limit we can still detect it across 76.5% of the samples taken from the real population.

## Summary

Across all three examples given in this paper, it is apparent that the most difficult and time consuming aspect of the Monte Carlo method for power is in creation of the MAHP. Great care must be taken to ensure that the generation of this hypothetical population is done properly. In situations where this approach may be necessary, in depth knowledge of the behavior of the data is necessary to ensure that the creation of the MAHP, the sampling from it, and the calculation of power is done correctly.

In all, the method is quite simple, aside from the creation of the MAHP. Create the hypothetical population to have the minimal difference determined, and incorporate the uncertainty. Generate the MAHP and consider many, say 1,000, repeated realizations from this distribution. In each situations the same statistical test is run. The proportion of times the test detects the difference incorporated into the MAHP, is the power.

## **Bibliography**

Hurwitz, A., Kitto, W., Remund, T., and Brownlow, J. 2011. “Estimation of location error for targeting using parametric, Monte Carlo and Bayesian techniques”. The ITEA Journal, 32(3): in press.

Ramsey, F., Schafer, D. 2001. *The Statistical Sleuth: A Course in Methods of Data Analysis 2ed.* Duxbury, Pacific Grove, CA.